

Elastic Network Models: Theoretical and Empirical Foundations

Yves-Henri Sanejouand

Abstract Fifteen years ago Monique Tirion showed that the low-frequency normal modes of a protein are not significantly altered when non-bonded interactions are replaced by Hookean springs, for all atom pairs whose distance is smaller than a given cutoff value. Since then, it has been shown that coarse-grained versions of Tirion's model are able to provide fair insights on many dynamical properties of biological macromolecules. In this chapter, theoretical tools required for studying these so-called Elastic Network Models are described, focusing on practical issues and, in particular, on possible artifacts. Then, an overview of some typical results that have been obtained by studying such models is given.

Key words: Protein, Normal Mode Analysis, Anisotropic Network Model, Gaussian Network Model, Low-frequency Modes, B-factors, Thermal Motion, Conformational Change, Functional Motion.

Yves-Henri Sanejouand
CNRS-UMR 6204, Faculté des Sciences, Nantes. e-mail:
yves-henri.sanejouand@univ-nantes.fr

1 Introduction

In 1996, Monique Tirion showed that the low-frequency normal modes of a protein (see section 3.1) are not significantly altered when Lennard-Jones and electrostatic interactions are replaced by Hookean (harmonic) springs, for all atom pairs whose distance is smaller than a given cutoff value [1]. In the case of biological macromolecules, this seminal work happened to be the first study of an Elastic Network Model (ENM). The ENM considered was an all-atom one, chemical bonds and angles being kept fixed through the use of internal coordinates, as often done in previous standard normal mode studies of proteins [2, 3, 4].

Soon afterwards, several coarse-grained versions of Tirion's ENM were proposed, in which each protein amino-acid residue is usually represented as a single bead and where most, if not all, chemical "details" are disregarded [5, 6], including atom types and amino-acid masses.

Since then, it has been shown that such highly simplified protein models are able to provide fair insights on the dynamical properties of biological macromolecules [5, 7, 8, 9], including those involved in their largest amplitude functional motions [10, 11], even in the case of large assemblies like RNA polymerase II [12], transmembrane channels [13, 14], whole virus capsids [15] or even the ribosome [16]. As a consequence, numerous applications have been proposed, noteworthy for exploiting fiber diffraction data [17], solving difficult molecular replacement problems [18, 19], or for fitting atomic structures into low-resolution electron density maps [19, 20, 21, 22, 23].

However, the idea that simple models can prove enough for capturing major properties of objects as complex as proteins had been put forward well before Tirion's introduction of ENMs in the realm of molecular biophysics. In the following, after a brief account of previous results supporting this claim (section 2), theoretical

tools required for studying an ENM are described (section 3), focusing on practical issues and, in particular, on possible artifacts. Then, an overview of typical results that have been obtained by studying protein ENMs is given (section 4).

2 Background

Indeed, coarse-grained models of proteins had been considered twenty years before M. Tirion's work, for studying what may well be the most complex phenomenon known at the molecular scale, namely, protein folding. Indeed, as soon as 1975, Michael Levitt and Arie Warshel proposed to model a protein as a chain of beads, each bead corresponding to the C_α atom of an amino-acid residue, the centroid of each amino-acid sidechain being taken into account with another bead grafted onto the chain [24]. That same year, Nobuhiro Go and his collaborators proposed an even simpler model in which the chain of beads is mounted on a two-dimensional lattice, each bead corresponding either to a single residue or, more likely, to a secondary structure element (e.g., an α -helix) of a protein [25]. Moreover, while the Levitt-Warshel model had been designed so as to study a specific protein, that is, a polypeptidic chain with a given sequence of amino-acid residues, the Go model focuses on the conformation of the chain, more precisely, on the set of pairs of amino-acids that are interacting together in the chosen (native) structure.

So, it is fair to view protein ENMs as off-lattice versions of the Go model.

Lattice models of proteins have been studied extensively since then so as to gain, for instance, a better understanding of the sequence-structure relationship. Noteworthy, if the chain is short enough, all possible conformations on the lattice can be enumerated, allowing for accurate calculations of thermodynamic quantities and

univoqual determination of the free energy minimum. Moreover, if the number of different amino-acids is small enough, then the whole sequence space can also be addressed. For instance, in the case of the tridimensional cubic lattice, a 27-mer chain has 103346 self-avoiding compact (i.e. cubic) conformations [26]. On the other hand, if only two kinds of amino-acids are retained, that is, if only their hydrophobic or hydrophilic nature is assumed to be relevant for the understanding of protein stability, then a 27-mer has 2^{27} different possible sequences. This is a large number, but it remains small enough so that for each sequence the lowest-energy compact conformation can be determined and, when a nearly-additive interaction energy is considered [27], the conclusion of such a systematic study happens to be an amazing one. Indeed, it was found that a few conformations (1% of them) are "preferred" by large sets of sequences [28]. Moreover, although each of these sets forms a neutral net in the sequence space, it is often possible to "jump" from a preferred conformation to another, as a consequence of single-point mutations [29].

While the former property is indeed expected to be a protein-like one, allowing to understand why proteins are able to accomodate so many different single-point mutations without significant loss of both their structure and function, it is only during the last few years that the latter one has been exhibited. In particular, using sequence design techniques, a pair of proteins with 95% sequence identity, but different folds and functions, was recently obtained [30]. If generic enough, such a property would help to understand how the various protein folds nowadays found on earth may have been "discovered" during the earliest phases of life evolution (e.g. prebiotic ones), since discovering a first fold could have proved enough for having access to many other ones, a single-point mutation after another.

In any case, this example shows how the study of simple models can help to think about, and maybe to understand better, major protein properties, in

particular because such models can be studied on a much larger scale than actual proteins.

3 Theoretical foundations

The vast majority of protein ENM studies rely on Normal Mode Analysis (NMA) [9]. Moreover, the hypotheses underlying this kind of analysis probably inspired the design of the first ENM. Actually, in her seminal work, M. Tirion performed NMA in order to show that similar results can be obtained by studying an ENM or a protein described at a standard, semi-empirical, level [1]. So, hereafter, the principles of NMA are briefly recalled (more details can be found in classic textbooks [31, 32]). Next, the close relationship between NMA and the different types of ENMs is underlined.

3.1 Normal Mode Analysis

Newton's equations of motion for a set of N atoms can not be solved analytically when N is large (namely, $N > 2$), except in rare instances like the following, rather general, one. Indeed, for small enough displacements of the atoms in the vicinity of their equilibrium positions, V , the potential energy of the studied system, can be approximated by the first terms of a Taylor series:

$$V = V_0 + \sum_{i=1}^{3N} \left(\frac{\partial V}{\partial r_i} \right)_0 (r_i - r_i^0) + \frac{1}{2} \sum_{i=1}^{3N} \sum_{j=1}^{3N} \left(\frac{\partial^2 V}{\partial r_i \partial r_j} \right)_0 (r_i - r_i^0)(r_j - r_j^0) \quad (1)$$

where r_i is the i^{th} coordinate, r_i^0 , its equilibrium value, and V_0 , the potential energy of the system at equilibrium.

Since, within the frame of classical physics, the exact value of V is meaningless (only potential energy differences are expected to play a physical role), V_0 can be zeroed. Moreover, since V_0 is a minimum of V , for each coordinate:

$$\left(\frac{\partial V}{\partial r_i}\right)_0 = 0$$

This yields:

$$V = \frac{1}{2} \sum_{i=1}^{3N} \sum_{j=1}^{3N} \left(\frac{\partial^2 V}{\partial r_i \partial r_j}\right)_0 (r_i - r_i^0)(r_j - r_j^0) \quad (2)$$

In other words, if the atomic displacements around an equilibrium configuration are small enough, then the potential energy of a system can be approximated by a quadratic form.

On the other hand, if the system is *not* under any constraint with an *explicit* time-dependence, then its kinetic energy can also be written as a quadratic form [31] and it is straightforward to show that, when both potential and kinetic energy functions are quadratic forms, then the equations of atomic motion have the following, analytical, solutions [31, 32, 33]:

$$r_i(t) = r_i^0 + \frac{1}{\sqrt{m_i}} \sum_{k=1}^{3N} C_k a_{ik} \cos(2\pi \nu_k t + \Phi_k) \quad (3)$$

where m_i is the atomic mass and where C_k and Φ_k , the amplitude and phasis of the so-called normal mode of vibration k , depend upon the initial conditions, that is, upon atomic positions and velocities at time $t = 0$. Noteworthy, C_k is a simple function of E_k , the total energy of mode k . In particular, if all modes have identical total energies, then:

$$C_k = \frac{\sqrt{2k_B T}}{2\pi \nu_k} \quad (4)$$

where T is the temperature and k_B the Boltzmann constant. This means that the amplitude of mode k goes as the inverse of its frequency, ν_k . As a matter of fact, when NMA is performed in the case of proteins, using standard all-atom force-fields, it can be shown that modes with frequencies below $30\text{-}100\text{ cm}^{-1}$ are responsible for 90-95% of the atomic displacements [34].

Note that such analytical solutions can provide various thermodynamic quantities like entropy, enthalpy, etc, and this, even at a quantum mechanical level of description [34].

In practice, the a_{ik} 's involved in eq. 3, which give the coordinate contributions to mode k , are obtained as the k^{th} eigenvector of \mathbf{H} , the mass-weighted Hessian of the potential energy, that is, the matrix whose element ij is:

$$\left(\frac{\partial^2 V}{\sqrt{m_i m_j} \partial r_i \partial r_j} \right)_0 \quad (5)$$

By definition, the $3N$ eigenvectors of a matrix like \mathbf{H} form an orthogonal basis set. This means that, when $k \neq l$:

$$\left(\frac{\partial^2 V}{\partial q_k \partial q_l} \right)_0 = 0$$

where q_k is the so-called normal coordinate, obtained by projecting the $3N$ mass-weighted cartesian coordinates onto eigenvector k , namely:

$$q_k = \sum_i^{3N} a_{ik} \sqrt{m_i} (r_i - r_i^0) \quad (6)$$

Moreover, the eigenvalues of \mathbf{H} , that is, the diagonal elements of the matrix obtained by expressing \mathbf{H} in this new basis set, provide the $3N$ frequencies of the system since, for each mode k :

$$\left(\frac{\partial^2 V}{\partial^2 q_k}\right)_0 = (2\pi\nu_k)^2$$

The eigenvalues and eigenvectors of a matrix are obtained by an operation called a diagonalization. In principle, for a real and symmetrical matrix like \mathbf{H} , such an operation is always possible. At a practical level, when the matrix size is not too large, that is, if the matrix can be stored in the computer memory, algorithms and methods available in standard mathematical packages allow to get its eigenvalues and eigenvectors at a CPU cost raising as nN^2 , where n is the number of requested eigensolutions. In other words, it is rather straightforward to obtain analytical solutions for the atomic motions, as long as small-amplitude displacements around a given, well-defined, equilibrium configuration are considered. Note that for a tridimensional system at equilibrium, at least six zero eigenvalues have to be obtained (except if the system is linear, in which case there are five of them), corresponding to the six possible rigid-body motions (translations or rotations) of the entire system. However, if the system is *not* at equilibrium, negative eigenvalues are usually observed. Moreover, significant mixing between rotation modes and some others can occur, leaving three zero eigenvalues only, that is, those corresponding to the three translation modes of the system [33].

The main drawback of NMA is obvious: the actual dynamics of a protein is much more complicated than assumed above. As a matter of fact, even on the short timescales considered within the frame of standard molecular dynamics simulations, a protein is able to jump from the attraction basin of an equilibrium configuration to another [35], and the number of these equilibrium configurations is so huge that it is unlikely for a nanosecond trajectory to visit one of them twice. In other words, while NMA focuses on protein dynamics at the level of a single minimum of the potential energy surface (PES), it is

well known that for proteins at room temperature the relevant PES is a highly complex, multi-minima, one.

NMA has several other drawbacks. For instance, starting from a given protein structure, e.g., as found in the Protein Databank (PDB), an equilibrium configuration has to be reached. This is usually done using energy-minimization techniques. As a consequence, the structure studied with NMA and a standard force field is always a distorted one, the C_α root-mean-square deviation (C_α -r.m.s.d) from the initial structure being typically of 1-2Å [9].

More importantly, within the frame of NMA, it is not obvious to take solvent effects into account, as the meaning of an equilibrium configuration in the case of an ensemble of molecules in the liquid state is unclear. As a matter of fact, the first NMA studies of proteins were performed *in vacuo* [2, 3, 4, 36]. Note that, nowadays, the availability of implicit solvent models, like EEF1 [37], offers a more satisfactory alternative.

However, as shown below, the main idea underlying the design of protein ENMs is not only to ignore the well-known drawbacks of NMA but, building upon its empirical successes, to add a few more on top of them.

3.2 The Elastic Network Model

In essence, there are two different types of ENMs, which differ by their dimensionality. The Gaussian Network Model (GNM), proposed by Ivet Bahar, Burak Erman and Turkan Haliloglu in 1997 [5, 38], is a one-dimension model while Tirion's model, later called the Anisotropic Network Model [39] (ANM), is a tridimensional one.

3.2.1 The Anisotropic Network Model

Although eq. 2 may look simple, it relies on a large number of parameters, namely, the elements of the Hessian matrix (eq. 5). In order to make it even simpler, M. Tirion proposed to replace eq. 2 by another quadratic form, namely:

$$V = \frac{1}{2}k_{enn} \sum_{d_{ij}^0 < R_c} (d_{ij} - d_{ij}^0)^2 \quad (7)$$

where d_{ij} is the actual distance between atoms i and j , d_{ij}^0 being their distance in the studied structure [1]. This amounts to set Hookean springs between all pairs of atoms less than R_c Ångströms away from each other. Note that in Tirion's work, as well as in most ANM studies (there are notable exceptions [40]), k_{enn} , the spring force constant, is the same for all atom pairs. When it is so, the role of k_{enn} is just to specify which system of units is used, R_c being the only physically relevant parameter of the model. In other words, when studying an ENM, the major drawback added with respect to standard NMA is that most atomic details are simply ignored.

However, considering eq. 7 instead of eq. 2 has several practical advantages. First, an energy minimization is not required any more, since the configuration whose energy is the absolute minimum one ($V = 0$) is known: it is the studied one. As a corollary, results obtained by studying ENMs are easier to reproduce. Indeed, an energy minimization not only introduces unwanted distortions in a structure, but it does it in a way that strongly depends upon the most tiny details of the protocole used, this, also as a consequence of the huge number of minima of a realistic PES for a biological macromolecule. Last but not least, as a straightforward consequence of eq. 7, the elements of the Hessian matrix (see eq. 5) are as simple as):

$$h_{ij} = -k_{enn} \frac{(x_i - x_j)(y_i - y_j)}{\sqrt{m_i m_j} d_{ij}^2} \quad (8)$$

where h_{ij} is the element corresponding to the x and y coordinates of atoms i and j .

3.2.2 The Gaussian Network Model

Because R_c , the cutoff value of an ANM, is usually rather small (see section 3.2.3), the corresponding Hessian matrix is sparse, that is, most of its elements (eq. 8) are zeroes. So, as proposed by I. Bahar, B. Erman and T. Haliloglu [5], it is tempting to go another step further into the simplification process and to consider the corresponding adjacency matrix, that is, the matrix whose elements are:

$$h_{ij} = -k_{enm} \quad (9)$$

when residues i and j are interacting ($h_{ij} = 0$ otherwise). Note that in the case of an adjacency matrix, as well as for the Hessian matrix of an ANM, h_{ii} , the diagonal element i , is so that:

$$h_{ii} = -\sum_{i \neq j} h_{ij} \quad (10)$$

Of course, with an adjacency matrix, information about directionality is missing. This is a major drawback of GNMs since this means that studying a GNM can only provide informations about motion amplitudes.

Note that GNMs are usually, if not always, set up at the residue level, while ANMs are sometimes studied at the atomic level, like in the seminal study of M. Tirion [1]. From now on, to underline such (not so common) cases, these latter models will be coined "all-atom ANMs".

3.2.3 The cutoff issue

The main, if not the only, parameter of an ENM is R_c . Although several studies have tried to justify the choice of a particular value for this parameter, typically by comparing calculated and experimental quantities, cutoff values over a wide range are still of common use, varying between 7 [41] and 16 Å [8].

For the most part, this probably reflects the fact the lowest-frequency modes of an ENM are usually "robust" [42], that is, little sensitive to the way the model is built. However, it is obvious that to be meaningful the value of R_c has to be on the small side. Putting it to an extreme: in the case of a GNM (see section 3.2.2), if R_c is so large that the adjacency matrix is completely filled with non-zero elements, its eigenvalues and eigenvectors, apart from being degenerate, will only depend upon N , the size of the system, and not upon its topology or its shape. As a consequence, they can for sure not provide any useful information. On the other hand, if R_c is too small, then the network of interacting residues is split into sub-networks, either free to rotate with respect to another one (in the case of an ANM) or completely independent from each other (in both ANM and GNM cases). Such dynamical properties are certainly not among those expected for a macromolecule, and this is why, in ANM studies, the smallest cutoff values used are of the order of 8-10 Å [10, 12], that is, larger than the typical distance between two interacting amino-acid residues in a protein, namely 6-7 Å [43, 44].

In practice, choosing a too small value for R_c yields additional zero eigenvalues.

So, if more than one (for a GNM) or six (for an ANM) zero eigenvalues are obtained, then it is highly recommended to increase R_c . Note that GNMs allow for the use of smaller values of R_c (a value of 7.3 Å is often chosen [41]) since in the

case of a mono-dimensional model a single connection is enough for avoiding any free translation of a group of atoms with respect to another. As a consequence, when a GNM is built with C_α atoms picked from a single protein chain, that is, when all amino-acid residues are chemically bonded to each other through peptidic bonds, a value of R_c as low as 4Å (the typical distance between two consecutive C_α atoms) can be used.

At first sight, it may seem that problems with small cutoff values could be solved with a distance-dependant spring force constant, as early proposed by Konrad Hinzen [6]. However, it is clear that an exponential term, for instance, introduces a typical length which, when too small, yields similar artifacts. Indeed, in such a case, the additional free rigid-body motions obtained with a too small value for R_c are expected to be replaced by low-frequency motions involving the same too little-connected groups of atoms.

Note that with ENMs other kinds of spurious low-frequency motions can be observed. For instance, in crystal structures, protein N- and C-terminal ends are often found to extend away from the rest of the structure. As a consequence, large amplitude, usually meaningless, motions of these (almost) free ends can be found among the lowest-frequency modes. So, in order to obtain significant and clear-cut results, it is highly recommended to begin an ENM study by "cleaning" the studied structure, namely, by removing such free ends.

A similar kind of spurious low-frequency motion can be observed with all-atom ANMs, in which groups of little-connected atoms are involved, typically those at the end of long sidechains [45]. Note that an elegant way to cure such artifacts is to use the RTB approximation [46, 47], which allows to remove from the Hessian matrix all contributions associated to motions occurring inside each "block" the system is split into (RTB stands for Rotation-Translation of Blocks). In most cases, a block corresponds to a given amino-acid residue but, while atom-atom interactions are

taken into account when the atoms belong to different blocks, each block can also correspond to a whole protein subunit, allowing for the study of systems as large as entire virus capsids [15].

4 Empirical foundations

As illustrated above, ENMs and NMA are closely related. As a consequence, the theoretical foundations of ENMs are for the most part those of NMA. However, when applied to complex molecular systems, NMA is known to have obvious drawbacks (see section 3.1). So, if NMA is still widely performed it is because of its empirical, sometimes unexpected, successes. As recalled below, most of these successes can also be achieved by studying ENMs.

4.1 *B-factors*

From eq. 3 and eq. 4, it is straightforward to show that $\langle \Delta r_i^2 \rangle$, the fluctuation of coordinate i with respect to its equilibrium value, is so that:

$$\langle \Delta r_i^2 \rangle = \frac{k_B T}{m_i} \sum_{k=1}^{n_{nz}} \frac{a_{ik}^2}{4\pi^2 \nu_k^2} \quad (11)$$

n_{nz} being the number of non-zero frequency normal modes of the system, namely, $n_{nz} = N - 1$ when a GNM is considered and $n_{nz} = 3N - 6$ when it is an ANM. However, in practice, since such fluctuations scale as the inverse of ν_k , the k^{th} mode frequency, a sum over the lowest-frequency normal modes of the system is usually enough for obtaining a fair approximation [34].

On the other hand, B_i , the crystallographic Debye-Waller factor (the so-called isotropic B-factors) of atom i , is expected to be related to the fluctuations of its

atomic coordinates through:

$$B_i = \frac{8\pi^2}{3} \langle \Delta x_i^2 + \Delta y_i^2 + \Delta z_i^2 \rangle \quad (12)$$

Although other physical factors are involved, like crystal disorder or lattice phonons, as well as non-physical ones, like the number of water molecules included in the structure refinement process by crystallographers, significant correlations between B-factor values predicted using eq. 11-12 and experimentally obtained ones have been reported in numerous cases.

For instance, in a study of 30 protein GNMs ($R_c = 7.5\text{\AA}$), a mean value of 0.62 ± 0.13 for this correlation coefficient was found [7]. Interestingly, in the same study, 26 other proteins were considered, for which accurate relaxation measurements had been measured by NMR, and the mean correlation between the corresponding fluctuations and those obtained using eq. 11 was found to be significantly higher, namely, 0.76 ± 0.04 , a remarkable agreement with the experimental data being achieved in several cases, with a correlation coefficient over 0.9 for four of them [7]. Amazingly, ANMs do not perform significantly better. For instance, in a study of 83 proteins ($R_c = 16\text{\AA}$), a mean value for the correlation coefficient of 0.68 ± 0.11 between predicted and isotropic B-factors was obtained [8] while, using the all-atom ANM ($R_c = 5\text{\AA}$) implemented in the Elnémo webserver [14], which makes use of the RTB approximation [46, 47], a very similar value of 0.68 ± 0.13 was found [8].

Note that in both studies mentioned above, when eq. 11 was used, overall translations or rotations of the entire protein within the crystal cell were excluded from the calculation, while it is well known that such motions are able to provide by themselves good correlations with experimental values [48]. In other words, much

better correlations with experimental B-factors can be obtained by mixing NMA predictions with protein rigid-body motions, the latter accounting partly for crystal disorder, but mostly for the phonon modes of the whole crystal. Interestingly, these latter modes can be taken into account within the frame of ENM studies, simply by including all crystal cell symmetries in the model [49, 50, 51].

Of course, such significant correlations with experimental data can only be obtained because the amplitude of atomic thermal fluctuations scales as the inverse of mode frequencies (see eq. 11). Indeed, with crude models like ENMs, the actual high-frequency modes of a protein can not be predicted, because such modes strongly depend upon the chemical details of the structure, only a few neighboring atoms (e.g., covalently bonded ones) being involved in the highest-frequency modes. This does not mean, though, that the high-frequency modes of an ENM can not bring any useful information. Indeed, they correspond to local motions occurring within the parts of the structure whose density is the highest [38]. Moreover, it has been shown that such regions often ly nearby enzyme active sites [52, 53].

On the other hand, the B-factor values themselves can not directly be obtained by studying ENMs, since their average is proportional to k_{enm} . Indeed, it is customary to choose k_{enm} so as to match average experimental B-factor values [9]. Another common way is to choose k_{enm} so as to reproduce the lowest-frequency of the system, as obtained using all-atom force-fields [52].

4.2 *The relationship with protein functional motions*

The seminal paper of M. Tirion ends with the statement that [1]:

Tests performed on a periplasmic maltodextrin binding protein (MBP) indicate that the slowest modes do indeed closely map the open form into the closed form (Tirion, in preparation).

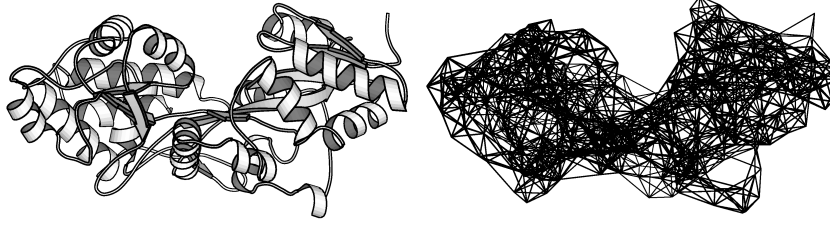


Fig. 1 Left: the open (ligand-free) form of maltodextrin binding protein (PDB identifier 1OMP). Right: the corresponding Elastic Network Model. Pairs of C_α atoms are linked by springs (plain lines) when they are less than 8Å from each other. Drawn with Molscrip [54].

The next paper of M. Tirion never came out but her result was confirmed a few years later, as part of a study of 20 protein ENMs ($R_c = 8\text{\AA}$) in both their ligand-free (open) and ligand-bound (closed) forms [10]. Indeed, for MBP, it was found that the overlap between its second lowest-frequency mode and its functional conformational change is close to 0.9. This means that 80% of the functional motion of MBP can be described by varying the normal coordinate associated to a single of its modes. Indeed, O_k , the overlap with mode k , is given by:

$$O_k = \frac{\sum_i \Delta r_i a_{ik}}{\sqrt{\sum \Delta r_i^2}} \quad (13)$$

where Δr_i is the variation of coordinate i between the open and the closed form after both structures have been superimposed [55]. On the other hand, since the modes of MBP form an orthogonal basis set, the following property holds:

$$\sum_{k=1}^{n_{nz}} O_k^2 = 1 \quad (14)$$

More generally, it was found that when the conformational change of a protein upon ligand binding happens to be highly collective, one of its low-frequency normal modes often compares well with the experimental motion (overlap over

0.5 [10]). Since then, a study of nearly 4,000 cases has confirmed this result [11], while another study of a set of proteins with similar functions and shapes, but various folds, namely DNA-dependant polymerases [12], has shown that the low-frequency modes of a protein, and hence the nature of its large amplitude motions, are likely to be determined by its shape [10, 56, 57].

Indeed, this latter point has recently been confirmed in a rather direct way, by considering ENMs built in such a way that each amino-acid interacts with a given number of neighbors (the closest ones). Then, at variance with cutoff-based ENMs, the rigidity of the system is fairly constant from a site to another. However, the relationship between the lowest-frequency modes of a protein and its functional motion is preserved. Specifically, it was found that the subspace defined by up to the 10-12 lowest-frequency modes of a protein is conserved, whatever model is used. Moreover, when no such, so-called robust, subspace exists, the functional motion of the protein is found to be either localized and/or of small amplitude (typically: less than $2\text{-}3\text{\AA}$ of C_α -r.m.s.d) [42].

In retrospect, these results make sense. First, a strong relationship between low-frequency modes and protein functional motions was first observed within the frame of NMA studies performed at a highly detailed, atomic level of description, noteworthy in the cases of lysozyme [58], hexokinase [59], citrate synthase [55] and hemoglobin [60]. Since, as recalled above, it was later found that such a relationship also holds when most chemical details are removed, it is clear that the property captured by NMA has to be a very general one. On the other hand, K. Hinsen has convincingly shown that the low-frequency modes of a protein can be used to split its structure into well-defined domains [6], with the additional advantage of a smooth, almost continuous, description of their boundaries. So, since it is well known that most large amplitude protein functional motions can be well described as combi-

nations of almost rigid-body motions of entire structural domains [61, 62], the relationship found between these motions and the low-frequency modes of ENMs is just another demonstration that whole quasi-rigid domain motions are involved in such modes. On the other hand, it is not that difficult to admit that the spatial clustering of amino-acids into domains can be revealed by studying protein dynamical properties, even at a crude level of description. A corollary of this line of thought is that ENMs should perform better, as far as low-frequency and large amplitude motions are concerned, in the case of large, multi-domain systems.

4.3 Applications

As illustrated above, NMA of ENMs seems to have a clear predictive power. So, given both the simplicity of these models and their coarse-grained nature, many applications have been proposed. For instance, as early suggested, being able to guess the pattern of atomic fluctuations through eq. 11 may prove useful for refining crystal structures [63, 64].

However, most applications take advantage of the possibility to predict atomic displacements through the reciproqual of eq. 6, namely:

$$r_i = r_i^0 + \frac{1}{\sqrt{m_i}} \sum_k^{n_{sub}} a_{ik} q_k \quad (15)$$

where n_{sub} is the number of low-frequency modes considered to be enough for performing an accurate prediction. In the simplest case, mode amplitudes can be varied arbitrarily, one mode after the other. Indeed, in the light of enough experimental data, the analysis of such trajectories can prove enough for getting insights about the nature of the functional motion of a protein [13, 65]. Some of the conformations thus obtained can also allow for solving difficult molecular-replacement problems,

although it is often necessary to explore at least a couple of modes in order to reach a useful conformation [18]. More generally, eq. 15 can be used so as to reduce the dimensionality of the system and, thus, to find more easily protein conformations fulfilling a given set of constraints. For instance, it has been used for fitting known structures into low-resolution electron density maps [19, 20, 21, 23] providing, for instance, more detailed structural data for systems of major interest, like the ribosome [22].

Note that eq. 15 is linear. As a consequence, atom motions follow straight lines and local distortions (of most chemical bonds, valence angles, etc) can not be avoided. So, for many applications, as well as for obtaining well-behaved normal mode trajectories, the conformations thus generated need to be "regularized" [18], using for instance a detailed all-atom force-field and standard energy-minimization techniques.

5 Conclusion

Fifteen years after their introduction in the realm of molecular biophysics [1], thanks to their simplicity as well as to their coarse-grained nature, Elastic Network Models are becoming more and more popular. Indeed, many applications have been proposed, noteworthy within the frame of various structural biology techniques.

From a theoretical point of view, their relationship with Normal Mode Analysis is obvious, since both approaches rely on a quadratic form for the energy function, the former, *par définition*, the latter, as a consequence of a small displacement, so-called harmonic (or linear) approximation.

From an empirical point of view, it has been extensively shown that normal mode studies of Elastic Network Models yield low-frequency, large amplitude and collective, motions which prove often similar to those obtained with an all-atom model and a standard empirical force-field.

This is likely to be a consequence of the robustness of these motions [42]. Moreover, such motions often provide fair predictions for the pattern of thermal atomic fluctuations (e.g. the crystallographic B-factors) or for the kind of functional motion a given protein can perform (e.g. its conformational change upon ligand binding).

References

1. Tirion, M. (1996). Low-amplitude elastic motions in proteins from a single-parameter atomic analysis. *Phys. Rev. Lett.* **77**, 1905–1908.
2. Noguti, T. & Go, N. (1982). Collective variable description of small-amplitude conformational fluctuations in a globular protein. *Nature* **296**, 776–778.
3. Go, N., Noguti, T. & Nishikawa, T. (1983). Dynamics of a small globular protein in terms of low-frequency vibrational modes. *Proc. Natl. Acad. Sci. USA* **80**, 3696–3700.
4. Levitt, M., Sander, C. & Stern, P. (1983). Normal-mode dynamics of a protein: Bovine pancreatic trypsin inhibitor. *Int. J. Quant. Chem.* **10**, 181–199.
5. Bahar, I., Atilgan, A. R. & Erman, B. (1997). Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold. Des.* **2**, 173–181.
6. Hinsen, K. (1998). Analysis of domain motions by approximate normal mode calculations. *Proteins* **33**, 417–429.
7. Micheletti, C., Lattanzi, G. & Maritan, A. (2002). Elastic properties of proteins: Insight on the folding process and evolutionary selection of native structures. *J. Mol. Biol.* **321**, 909–921.
8. Kondrashov, D., Van Wynsberghe, A., Bannen, R., Cui, Q. & Phillips, G. (2007). Protein structural variation in computational models and crystallographic data. *Structure* **15**, 169–177.
9. Bahar, I. & Cui, Q., eds. (2005). *Normal Mode Analysis: Theory and Applications to Biological and Chemical Systems. C&H/CRC Mathematical & Computational Biology Series, vol. 9.*

- CRC press, Boca Raton.
10. Tama, F. & Sanejouand, Y. H. (2001). Conformational change of proteins arising from normal mode calculations. *Prot. Engineering* **14**, 1–6.
 11. Krebs, W. G., Alexandrov, V., Wilson, C. A., Echols, N., Yu, H. & Gerstein, M. (2002). Normal mode analysis of macromolecular motions in a database framework: Developing mode concentration as a useful classifying statistic. *Proteins* **48**, 682–695.
 12. Delarue, M. & Sanejouand, Y.-H. (2002). Simplified normal modes analysis of conformational transitions in DNA-dependant polymerases: the Elastic Network Model. *J. Mol. Biol.* **320**, 1011–1024.
 13. Valadie, H., Lacapere, J.-J., Sanejouand, Y.-H. & Etchebest, C. (2003). Dynamical properties of the MscL of Escherichia coli: a Normal Mode Analysis. *J. Mol. Biol.* **332**, 657–674.
 14. Suhre, K. & Sanejouand, Y.-H. (2004). Elnémo: a normal mode server for protein movement analysis and the generation of templates for molecular replacement. *Nucl. Ac. Res.* **32**, W610–W614.
 15. Tama, F. & Brooks III, C. (2002). The mechanism and pathway of pH induced swelling in cowpea chlorotic mottle virus. *J. Mol. Biol.* **318**, 733–747.
 16. Tama, F., Valle, M., Frank, J. & Brooks III, C. L. (2003). Dynamic reorganization of the functionally active ribosome explored by normal mode analysis and cryo-electron microscopy. *Proc. Natl. Acad. Sci. USA* **100**, 9319–9323.
 17. Tirion, M., ben Avraham, D., Lorenz, M. & Holmes, K. (1995). Normal modes as refinement parameters for the F-actin model. *Biophys. J.* **68**, 5–12.
 18. Suhre, K. & Sanejouand, Y.-H. (2004). On the potential of normal mode analysis for solving difficult molecular replacement problems. *Act. Cryst. D* **60**, 796–799.
 19. Delarue, M. & Dumas, P. (2004). On the use of low-frequency normal modes to enforce collective movements in refining macromolecular structural models. *Proc. Natl. Acad. Sci. USA* **101**, 6957–6962.
 20. Tama, F., Miyashita, O. & Brooks III, C. L. (2004). Flexible multi-scale fitting of atomic structures into low-resolution electron density maps with elastic network normal mode analysis. *J. Mol. Biol.* **337**, 985–999.
 21. Hinsen, K., Reuter, N., Navaza, J., Stokes, D. L. & Lacapere, J. J. (2005). Normal mode-base fitting of atomic structure into electron density maps: Application to sarcoplasmic reticulum Ca-ATPase. *Biophys. J.* **88**, 818–827.

22. Mitra, K., Schaffitzel, C., Shaikh, T., Tama, F., Jenni, S., Brooks 3rd, C., Ban, N. & Frank, J. (2005). Structure of the E. coli protein-conducting channel bound to a translating ribosome. *Nature* **438**, 318–324.
23. Suhre, K., Navaza, J. & Sanejouand, Y.-H. (2006). Norma: a tool for flexible fitting of high resolution protein structures into low resolution electron microscopy derived density maps. *Act. Cryst. D* **62**, 1098–1100.
24. Levitt, M. & Warshel, A. (1975). Computer simulation of protein folding. *Nature* **253**, 694–698.
25. Taketomi, H., Ueda, Y. & Go, N. (1975). Studies on protein folding, unfolding and fluctuations by computer simulation. I. The effect of specific amino acid sequence represented by specific inter-unit interactions. *Int. J. Pept. Prot. Res.* **7**, 445–459.
26. Shakhnovich, E. & Gutin, A. (1990). Enumeration of all compact conformations of copolymers with random sequence of links. *J. Chem. Phys.* **93**, 5967.
27. Li, H., Tang, C. & Wingreen, N. (1997). Nature of driving force for protein folding: A result from analyzing the statistical potential. *Phys. Rev. Letters* **79**, 765–768.
28. Li, H., Helling, R., Tang, C. & Wingreen, N. (1996). Emergence of preferred structures in a simple model of protein folding. *Science* **273**, 666–669.
29. Trinquier, G. & Sanejouand, Y. H. (1999). New protein-like properties of cubic lattice models. *Phys. Rev. E* **59**, 942–946.
30. He, Y., Chen, Y., Alexander, P., Bryan, P. & Orban, J. (2008). NMR structures of two designed proteins with high sequence identity but different fold and function. *Proc. Natl. Acad. Sci. USA* **105**, 14412.
31. Goldstein, H. (1950). *Classical Mechanics*. Addison-Wesley, Reading, MA.
32. Wilson, E., Decius, J. & Cross, P. (1955). *Cross. Molecular Vibrations*. McGraw-Hill, New York.
33. Sanejouand, Y.-H. (1990). *Ph. D. Thesis*. Université de Paris XI, Orsay, France.
34. Levy, R., Perahia, D. & Karplus, M. (1982). Molecular dynamics of an alpha-helical polypeptide: temperature dependence and deviation from harmonic behavior. *Proc. Natl. Acad. Sci. USA* **79**, 1346–1350.
35. Elber, R. & Karplus, M. (1987). Multiple conformational states of proteins: A molecular dynamics analysis of myoglobin. *Science* **235**, 318–321.
36. Brooks, B. & Karplus, M. (1983). Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor. *Proc. Natl. Acad. Sci. USA* **80**, 6571–6575.

37. Lazaridis, T. & Karplus, M. (1999). Effective energy function for proteins in solution. *Proteins* **35**, 133–152.
38. Haliloglu, T., Bahar, I. & Erman, B. (1997). Gaussian dynamics of folded proteins. *Phys. Rev. Letters* **79**, 3090–3093.
39. Atilgan, A., Durell, S., Jernigan, R., Demirel, M., Keskin, O. & Bahar, I. (2001). Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys. J.* **80**, 505–515.
40. Hinsen, K. & Kneller, G. (1999). A simplified force field for describing vibrational protein dynamics over the whole frequency range. *J. Chem. Phys.* **111**, 10766.
41. Kundu, S., Melton, J., Sorensen, D. & Phillips Jr, G. (2002). Dynamics of proteins in crystals: comparison of experiment with simple models. *Biophysical journal* **83**, 723–732.
42. Nicolay, S. & Sanejouand, Y.-H. (2006). Functional modes of proteins are among the most robust. *Phys. Rev. Lett.* **96**, 078104.
43. Miyazawa, S. & Jernigan, R. (1985). Estimation of effective inter-residue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules* **18**, 534–552.
44. Miyazawa, S. & Jernigan, R. (1996). Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term for simulation and threading. *J. Mol. Biol.* **256**, 623–644.
45. Tama, F. (2000). *Ph. D. Thesis*. Université Paul Sabatier, Toulouse, France.
46. Durand, P., Trinquier, G. & Sanejouand, Y. H. (1994). A new approach for determining low-frequency normal modes in macromolecules. *Biopolymers* **34**, 759–771.
47. Tama, F., Gadea, F.-X., Marques, O. & Sanejouand, Y.-H. (2000). Building-block approach for determining low-frequency normal modes of macromolecules. *Proteins* **41**, 1–7.
48. Kuriyan, J. & Weis, W. (1991). Rigid protein motion as a model for crystallographic temperature factors. *Proc. Natl. Acad. Sci. USA* **88**, 2773.
49. Simonson, T. & Perahia, D. (1992). Normal modes of symmetric protein assemblies. Application to the tobacco mosaic virus protein disk. *Biophys. J.* **61**, 410–427.
50. Hinsen, K. (2008). Structural flexibility in proteins: Impact of the crystal environment. *Bioinformatics* **24**, 521.
51. Riccardi, D., Cui, Q. & Phillips Jr, G. (2009). Application of elastic network models to proteins in the crystalline state. *Biophys. J.* **96**, 464–475.
52. Juanico, B., Sanejouand, Y.-H., Piazza, F. & De Los Rios, P. (2007). Discrete breathers in nonlinear network models of proteins. *Phys. Rev. Lett.* **99**, 238104.

53. Sacquin-Mora, S., Laforet, E. & Lavery, R. (2007). Locating the active sites of enzymes using mechanical properties. *Proteins* **67**, 350–359.
54. Kraulis, P. (1991). Molscript: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Cryst.* **24**, 946–950.
55. Marques, O. & Sanejouand, Y.-H. (1995). Hinge-bending motion in citrate synthase arising from normal mode calculations. *Proteins* **23**, 557–560.
56. Lu, M. & Ma, J. (2005). The Role of Shape in Determining Molecular Motions. *Biophys. J.* **89**, 2395–2401.
57. Tama, F. & Brooks, C. (2006). Symmetry, form, and shape: Guiding principles for robustness in macromolecular machines. *Annu Rev Biophys Biomol Struct* **35**, 115–33.
58. McCammon, J. A., Gelin, B. R., Karplus, M. & Wolynes, P. (1976). The hinge-bending mode in lysozyme. *Nature* **262**, 325–326.
59. Harrison, R. (1984). Variational calculation of the normal modes of a large macromolecules: Methods and some initial results. *Biopolymers* **23**, 2943–2949.
60. Perahia, D. & Mouawad, L. (1995). Computation of low-frequency normal modes in macromolecules: improvements to the method of diagonalization in a mixed basis and application to hemoglobin. *Comput. Chem.* **19**, 241–246.
61. Branden, C., Tooze, J. et al. (1991). *Introduction to protein structure*. Garland Publishing New York.
62. Gerstein, M. & Krebs, W. (1998). A database of macromolecular motions. *Nucl. Acid. Res.* **26**, 4280–4290.
63. Diamond, R. (1990). On the use of normal modes in thermal parameter refinement: theory and application to the bovine pancreatic trypsin inhibitor. *Acta Cryst. A* **46**, 425–435.
64. Kidera, A. & Go, N. (1990). Refinement of Protein Dynamic Structure: Normal Mode Refinement. *Proc. Natl. Acad. Sci. USA* **87**, 3718–3722.
65. Sanejouand, Y.-H. (1996). Normal-mode analysis suggests important flexibility between the two N-terminal domains of CD4 and supports the hypothesis of a conformational change in CD4 upon HIV binding. *Prot. Eng.* **9**, 671–677.